

METHOD FOR FORMING AN ONO STRUCTURE IN ONE CHAMBER

BACKGROUND OF THE INVENTION

5

1. Field of the Invention

The present invention generally relates to a method for forming an oxide-nitride-oxide structure of a memory cell, and more particularly to a method for forming an oxide-nitride-oxide structure in one chamber.

2. Description of the Prior Art

Flash memory is the most potential memory in the semiconductor industry. Flash memory has been broadly applied to repeatedly access data and to remain assessable during power break down, such as the film of digital camera or the basic input-output system of a motherboard. Because flash memory has the advantages of electrically erasable and programmable mechanisms, it can simultaneously proceed the erase and the program mechanisms to all flash memory cells in the whole memory array. Accordingly, how to advance the performance and reduce the cost of the flash memory became an important subject.

10002978-120601
The conventional flash memory structure is shown in FIG. 1A. First a substrate 100 is provided, and a field oxide layer 101 is formed on the substrate 100 by local oxidation (LOCOS) to define an active area. A tunnel oxide layer 102 is formed on the substrate 100. A floating gate 104 is formed on the tunnel oxide layer 102. The tunnel oxide layer 102 and the ONO structure are dielectric layers. The floating gate 104 is poly for storing and erasing data. A control gate 108 is formed over the floating gate 104. A dielectric layer 106 with an ONO structure comprising a sandwiched type of oxide-nitride-oxide layer 106 is formed between the control gate 108 and the floating gate 104. The ONO multi-layered dielectric containing three layers; namely a first oxide layer 106a, a nitride layer 106b, and a second oxide layer 106c. An N-type source/drain 110/112 region is formed in the substrate 100 beside the floating gate 104. Oxide spacers 114 are formed on the side-wall of the floating gate 104 and the control gate 108 to protect the transistor from being damaged.

The function of a flash memory operates when a positive voltages are applied to drain 112 and control gate 108 that force electrons to inject from the substrate near drain 112 to floating gate 104. To erase the memory device, a negative voltage is applied to the

control gate 108 and positive voltage is applied to source 110 that force electrons through tunnel to source region.

The traditional ONO structure of a flash memory is shown in FIG. 1B, first silane (SiH_4) and nitrogen oxides (N_2O) is introduced to deposit the bottom oxide layer 106A of the ONO structure, when the reaction is completed. In the next step, silane (SiH_4) and ammonia (NH_3) are introduced to deposit the nitride layer 106B. In the third step, silane (SiH_4) and nitrogen oxides (N_2O) are introduced again to deposit the upper oxide layer 106C. A complete ONO structure of a flash memory is thus formed.

The traditional fabrication method for producing ONO structure 106 is conducted by LPCVD method under a temperature of about 400 to 500 °C in a batch type furnace process. The batch type processing is proceeded in a boat carrying about 150 pieces of wafers. The disadvantage of the batch typed processing is that if there is any accident occurs, for example, a short circuit occurrence would cause a large cost of loss.

In the traditional ONO layer fabrication, it is also noted that when the layers of oxide and nitride are constructed, it would be necessary that the process must be conducted by separate chamber. A separate chamber for proceeding the construction of an ONO

structure is a time consuming task especially when the process is proceeded under a thermal treatment, the annealing process associated with the thermal treatment requires a long time to re-organize the structure of the ONO structure.

Another defect associated with the structure of ONO is due to the existence of rough interfaces among oxide-nitride-oxide layers, especially when there is a higher nitride film roughness on this ONO structure. This roughness contrast can be an obstacle for the promotion of flash memory speed efficiency.

In view of the prior art described, it is a desire to provide a method for forming an oxide-nitride-oxide structure in one chamber to improve nitride layer roughness problem in the ONO structure.

SUMMARY OF THE INVENTION

In accordance with the present invention, a method for forming an oxide-nitride-oxide structure in one chamber that substantially can use one chamber to form ONO structure in conventional process.

One object of the present invention is to provide a method for forming an oxide-nitride-oxide structure in one chamber to increase

flash memory speed efficiency.

Another object of the present invention is to provide a method for forming an oxide-nitride-oxide structure in one chamber to improve nitride layer roughness problem in the ONO structure.

In order to achieve the above object, the present invention provides a method for forming an oxide-nitride-oxide structure in one chamber. The method at least includes the following steps. First of all, a substrate is provided. Then, a first oxide layer is formed on the substrate. Next, a first buffer layer is formed on the first oxide layer, and a silicon nitride layer is formed on the first buffer layer. Next, a second buffer layer is formed on the silicon nitride layer. Finally, a second oxide layer is formed on the second buffer layer.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing aspects and many of the attendant advantages of this invention will become more readily appreciated as the same becomes better understood by referring to the following detailed description, when taken in conjunction with the accompanying drawings, wherein:

FIG. 1A is a diagram showing the flash memory having an

ONO structure in prior art techniques;

FIG. 1B is a diagram of the ONO structure in prior art techniques;

FIG.2A is cross-sectional views of a method for forming an oxide-nitride-oxide structure in one chamber in accordance with one preferred embodiment of the present invention;

FIG. 2B is a diagram of the ONO structure in this invention technique.

DESCRIPTION OF THE PREFERRED EMBODIMENT

Some sample embodiments of the invention will now be described in greater detail. Nevertheless, it should be recognized that the present invention can be practiced in a wide range of other embodiments besides those explicitly described, and the scope of the present invention is expressly not limited except as specified in the accompanying claims.

Moreover, while the present invention is illustrated by a number of preferred embodiments directed to silicon semiconductor devices, it is not intended that these illustrations be a limitation on the scope or applicability of the present invention. Thus, it is not

intended that the semiconductor devices of the present invention be limited to the structures illustrated. These devices are included to demonstrate the utility and application of the present invention to presently preferred embodiments.

5

Further, various parts of the semiconductor elements have not been drawn to scale. Certain dimensions have been exaggerated in relation to other dimensions in order to provide a clear illustration and understanding of the present invention. For example, although the embodiments illustrated herein are shown in two dimensional views with various regions having width and depth, it should be clearly understood that these regions are illustrations of only a portion of a single cell of a device, which may include a plurality of such cells arranged in a three-dimensional structure. Accordingly, these regions will have three dimensions, including length, width and depth, when fabricated in an actual device.

10

15

FIG.2A is cross-sectional views of a method for forming an oxide-nitride-oxide structure in one chamber in accordance with one preferred embodiment of the present invention;

20

FIG. 2B is a diagram of the ONO structure in this invention technique.

The invention flash memory structure is shown in FIG.2A. Firstly, a semiconductor substrate 200 is provided, and a field oxide layer 201 is formed on the substrate 200 by local oxidation (LOCOS) to define an active area. The semiconductor substrate 200 comprises silicon. A tunnel oxide layer 202 is formed on the semiconductor substrate 200. A floating gate 204 is formed on the tunnel oxide layer 202. The tunnel oxide layer 202 and the ONO structure are dielectric layers. The floating gate 204 is poly for storing and erasing data. A control gate 208 is formed over the floating gate 204. A dielectric layer 206 with an O-SiON-N-SiON-O structure comprising a sandwiched type of oxide-nitride-oxide layer 206 is formed between the control gate 208 and the floating gate 204. The O-SiON-N-SiON-O multi-layered dielectric containing five layers; namely a first oxide layer 206a, a first buffer layer 206b, a nitride layer 206c, a second buffer layer 206d, and a second oxide layer 206e. An N-type source/drain 210/212 region is formed in the substrate 200 beside the floating gate 204. Oxide spacer 214 are formed on the side-wall of the floating gate 204 and the control gate 208 to protect the transistor from being damaged.

The function of a flash memory operates when a positive voltages are applied to drain 212 and control gate 208 that force electrons to inject from the substrate near drain 212 to floating gate

204. To erase the memory device, a negative voltage is applied to the control gate 208 and positive voltage is applied to source 210 that force electrons to tunnel source region.

5 The invention O-SiON-N-SiON-O structure 206 of a flash memory is shown in FIG.2B. The O-SiON-N-SiON-O structure 206 includes a first oxide layer 206a overlying the floating gate 204. A first buffer layer 206b overlies first oxide layer 206a. A nitride layer 206c overlies first buffer layer 206b. A second buffer layer 206d overlies nitride layer 206c. A second oxide layer 206e overlies second buffer layer 206d. The O-SiON-N-SiON-O structure 206 is deposited in one wafer and in one chamber. The O-SiON-N-SiON-O structure 206 also can apply another device that have ONO structure such as NROM.

10 The first oxide layer 206a is formed to a thickness between about 30 and 100 angstroms on the floating gate 204 by introducing silane gas (SiH_4) at between about 5 and 200 seconds and nitrogen oxide gas (N_2O) at between about 5 and 200 seconds. The first oxide layer 206a deposition rate is between about 0.1 and 20 Å/sec. A flow of silane gas is between about 0.1 and 100 sccm. A flow of nitrogen oxide gas is between about 10 and 1000 sccm. The first oxide layer 206a is formed by using a low pressure vapor chemical deposition

method(LPCVD). The LPCVD process is carried out at a temperature of between about 500 °C and 2000 °C . In the embodiment, temperature of this layer is preferable 800 °C .

5 The first buffer layer 206b is formed to a thickness of between about 0.5 and 5 angstroms on the first oxide layer 206a by using a low pressure vapor chemical deposition (LPCVD) method and introducing silane gas (SiH_4) at between about 0.5 and 20 seconds, nitrogen oxide gas (N_2O) at between about 0.5 and 20
10 seconds, and ammonia (NH_3) gas at between about 0.5 and 20 seconds. The LPCVD process is carried out at a temperature of between about 500°C and 2000°C . In the embodiment, temperature of this layer is preferable 800 °C . The first buffer layer 206b deposition rate is between about 0.1 and 5 Å/sec. A flow of silane
15 gas is between about 0.1 and 1000 sccm. A flow of nitrogen oxide gas is between about 0.1 and 1000 sccm. A flow of ammonia gas is between about 0.1 and 1000 sccm. The first buffer layer 206b is silicon oxynitride (SiON) layer. The first buffer layer 206b is deposited by gradually reducing the supply of nitrogen oxide gas, in
20 the mean while, gradually increasing the supply of ammonia gas. In this step, silane gas is also provided for the deposition continually.

The nitride layer 206c is formed to a thickness of between

about 30 and 200 angstroms on first buffer layer 206b by using a low pressure vapor chemical deposition method (LPCVD) and introducing silane gas (SiH_4) at between about 10 and 300 seconds and ammonia (NH_3) gas at between about 10 and 300 seconds. The
5 LPCVD process is carried out at a temperature of between about 500 °C and 2000°C. In the embodiment, temperature of this layer is preferable 800 °C. The nitride layer 206c deposition rate is between about 0.5 and 100 Å/sec. A flow of silane gas is between about 0.1 and 100 sccm. A flow of ammonia gas is between about 10 and 1000
10 sccm. The nitride layer 206c is deposited by stop providing the nitrogen oxide gas and start providing the ammonia gas for deposition. In this step, silane gas is also provided for the deposition continually.

15 The second buffer layer 206d is formed to a thickness of between about 0.5 and 5 angstroms on the nitride layer 206c by using a low pressure vapor chemical deposition method and introducing silane gas (SiH_4) at between about 0.5 and 20 seconds, nitrogen oxide gas (N_2O) at between about 0.5 and 20 seconds, and
20 ammonia (NH_3) gas at between about 0.5 and 20 seconds. The LPCVD process is carried out at a temperature of between about 500 °C and 2000°C. In the embodiment, temperature of this layer is preferable 800 °C. The second buffer layer 206d deposition rate is

between about 0.1 and 5 Å/sec. A flow of silane gas is between about 0.1 and 100 sccm. A flow of nitrogen oxide gas is between about 0.1 and 1000 sccm. A flow of ammonia gas is between about 0.1 and 1000 sccm. The second buffer layer 206d is silicon oxynitride (SiON) layer. The second buffer layer 206d is deposited by gradually reducing the supply of ammonia gas, in the mean while, gradually increasing the supply of nitrogen oxide gas. In this step, silane gas is also provided for the deposition continually.

The second oxide layer 206e is formed to a thickness of between about 30 and 100 angstroms on second buffer layer 206d by using a low pressure vapor chemical deposition method and introducing silane gas (SiH₄) at between about 5 and 200 seconds and nitrogen oxide gas (N₂O) at between about 5 and 200 seconds. The LPCVD process is carried out at a temperature of between about 500°C and 2000°C. In the embodiment, temperature of this layer is preferable 800 °C. The second oxide layer 206e deposition rate is between about 0.1 and 20 Å/sec. A flow of silane gas is between about 0.1 and 100 sccm. A flow of nitrogen oxide gas is between about 10 and 1000 sccm. The second oxide layer 206e is deposited by stop providing ammonia gas and supply only the nitrogen oxide gas for deposition. In this step, silane gas is also provided for the deposition continually.

Thus, an O-SiON-N-SiON-O structure with smooth buffer layer in the intermediate layer is thus constructed. The layer between first oxide layer and nitride layer is the layer of silicon oxynitride (SiON) with a smooth buffer layer that is initially an oxide composition and has a tendency slowly turning to nitride until the composition is totally nitride. In the side of nitride layer, there is a layer of silicon oxynitride (SiON) with a smooth buffer layer that is initially a nitride composition and has a tendency slowly turning to oxide until the composition is totally oxide. This smooth buffer layer can be easily seen from reflective index changes of oxide layer with a value of 1.46 into nitride layer with a value of 2.0. It is obvious that the first buffer layer SiON of the constructed O-SiON-N-SiON-O structure has a reflective index changes varies slowly from 1.46 to 2.0 and also a reflective index of 2.0 slowly changes to 1.46 on the second buffer layer SiON of the constructed O-SiON-N-SiON-O structure.

The SiON lies between oxide to nitride layers and nitride to oxide layer is termed as a smooth buffer layers which is used to reduce the boundary roughness among ONO layers.

It should be emphasized here, the deposited O-SiON-N-

SiON-N layer is conducted in one chamber such that the time required in the annealing process associated with the various thermal treatments can be reduced a lot.

5 The foregoing description of the preferred embodiments in this invention has been presented for purpose of illustration and description. It is not intended to be exhaustive or to limit the invention to the precise form disclosed. Numerous modifications or variations are possible in light of the above teaching. The
10 embodiments were chosen and described to provide the best illustration of the principles of the invention and its practical application to thereby enable one of the ordinary skill in the art to utilize the invention in various embodiments and with various modifications which are suited to the particular use contemplated.
15 All such modifications and variations are within the scope of the invention as determined by appended claims when interpreted in accordance with the breadth to which they are legally and equitably entitled.

20 The method for forming an oxide-nitride-oxide structure in one chamber using the above explained method, has the following advantages:

1. The present invention is to provide a method for

forming an oxide-nitride-oxide structure in one chamber to increase flash memory speed efficiency.

2. The present invention is to provide a method for forming an oxide-nitride-oxide structure in one chamber to improve nitride layer roughness problem in the ONO structure.

While this invention has been described with reference to illustrative embodiments, this description is not intended to be construed in a limiting sense. Various modifications and combinations of the illustrative embodiments, as well as other embodiments of the invention, will be apparent to persons skilled in the art upon reference to the description. It is therefore intended that the appended claims encompass any such modifications or embodiments.